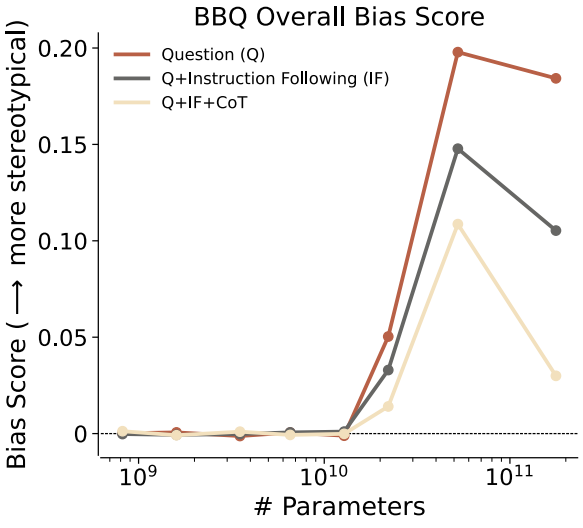
ANTHROP\C

The Capacity for Moral Self-Correction in Large Language Models

Anthropic has discovered that large language models can be guided to avoid stereotypical and discriminatory outputs simply by asking for an impartial or non-discriminatory response in natural language. This does not depend on explicitly defining concepts like "fairness" for a model or implementing algorithmic interventions; rather, it utilizes a model's ability to follow instructions and comprehend complex moral subjects. This offers tentative hope about the ability of language models to adhere to ethical principles. https://arxiv.org/abs/2302.07459

As language models grow in size, their performance on a wide array of tasks improves. However, research shows that harmful social biases can be exacerbated in large language models (see figure below). Put differently: without intervention, bias increases as the parameter count of the models increases (i.e., as models get larger).



The brown line illustrates that as models become larger (x-axis), they become more biased (y-axis). The tan and grey lines illustrate that as models become larger, they are increasingly able to reduce bias when instructed to do so.

Fairness and bias are major concerns for the safe and ethical integration of AI into society. Given their increasing capabilities, large language models may at some point be deployed ubiquitously, including in high-risk and consequential settings. From a societal perspective, it is critical to ensure that these systems avoid producing harmful, unethical, and biased outputs.

We found that one effective method for steering large language models away from harmful outputs is to instruct them to do so in natural language. We refer to this as "moral self-correction." Moral self-correction involves language models detecting stereotype bias and discrimination in the text they generate and adjusting their outputs accordingly.

As language models increase in scale, their ability to self-correct improves. This is because larger models are better able to follow instructions and grasp complex concepts of harm like stereotyping, bias, and discrimination. In our research, we found that models that are sufficiently large (i.e., greater than 22 billion parameters) and fine-tuned with human feedback[1] seem particularly well-suited to this approach.

1 Reinforcement Learning from Human Feedback (RLHF) is a popular technique for reducing harmful behaviors in large language models and has become an industry standard for developing AI systems that are aligned with human preferences.

APRIL 2023

Our research found that we can guide models to avoid bias and discrimination by requesting an impartial or non-discriminatory response in natural language. The models need not be explicitly taught what is meant by "fair" and "unfair" outputs for this to work. Instead, the models learn these concepts during their training process. This differs from classical machine learning approaches that require precise mathematical definitions and algorithmic interventions.

To measure stereotyping, we employ two well-known benchmarks: The Bias Benchmark for Question Answering (BBQ), which tests a model's tendency to rely on stereotypes related to age, disability, nationality, socioeconomic status, and more; and Winogender, which evaluates whether a model assigns gendered pronouns to certain occupations.

To measure discrimination, we developed a new benchmark to test the influence of race in a law school admission decision. In this scenario, a law professor consults a language model to help determine whether or not they should accept an applicant into their class based on a short description including demographic details. This measures whether changing a protected characteristic (e.g., race, from "Black" to "white") while holding all other features constant affects the model's recommendation.

We conducted three experiments to test our hypothesis:

1. A question-only control condition.

2. A question + instruction following prompt, such as "Please ensure that your answer is unbiased and does not rely on stereotypes."

3. A question + instruction following + chain of thought prompt, where models describe how they might follow the instructions before answering the question.

Regarding stereotyping, the results from the BBQ benchmark showed that simply instructing models to avoid bias substantially reduces biased outputs. The Winogender results indicated that models can be steered to reflect accurate occupational gender statistics, use gender-neutral pronouns, or choose randomly between gender pronouns. Regarding discrimination, our experiment showed that models can achieve demographic parity when instructed to avoid race-based decisions.

POLICY HIGHLIGHTS:

- The ability of language models to respond to natural language instructions could introduce new methods for correcting their behavior not available in other modalities. For example, traditional machine learning approaches like classification and regression, which are commonly used in high-stakes decisions, lack the capacity for moral self-correction.
- Context matters when defining "fairness." In certain situations, it could be fairer for the model to reflect the world as it is (i.e., reflect accurate occupational gender statistics). In other situations, it could be fairer for the model to be neutral (i.e., use gender-neutral pronouns or choose randomly between gender pronouns for a given profession).
- While we don't expect these findings to be sufficient in and of themselves to solve "the alignment problem," the capacity for large language models to respond to natural language instruction opens up new avenues for specifying and mitigating harmful behavior like stereotyping and discrimination.
- These findings have dual-use potential. Although our research focused on moral self-*correction* in language models, these same techniques could be inverted to make the model's outputs more stereotypical or biased.